JUSTC

Article

http://justc.ustc.edu.cn

Received: July 20, 2022; Accepted: September 05, 2022

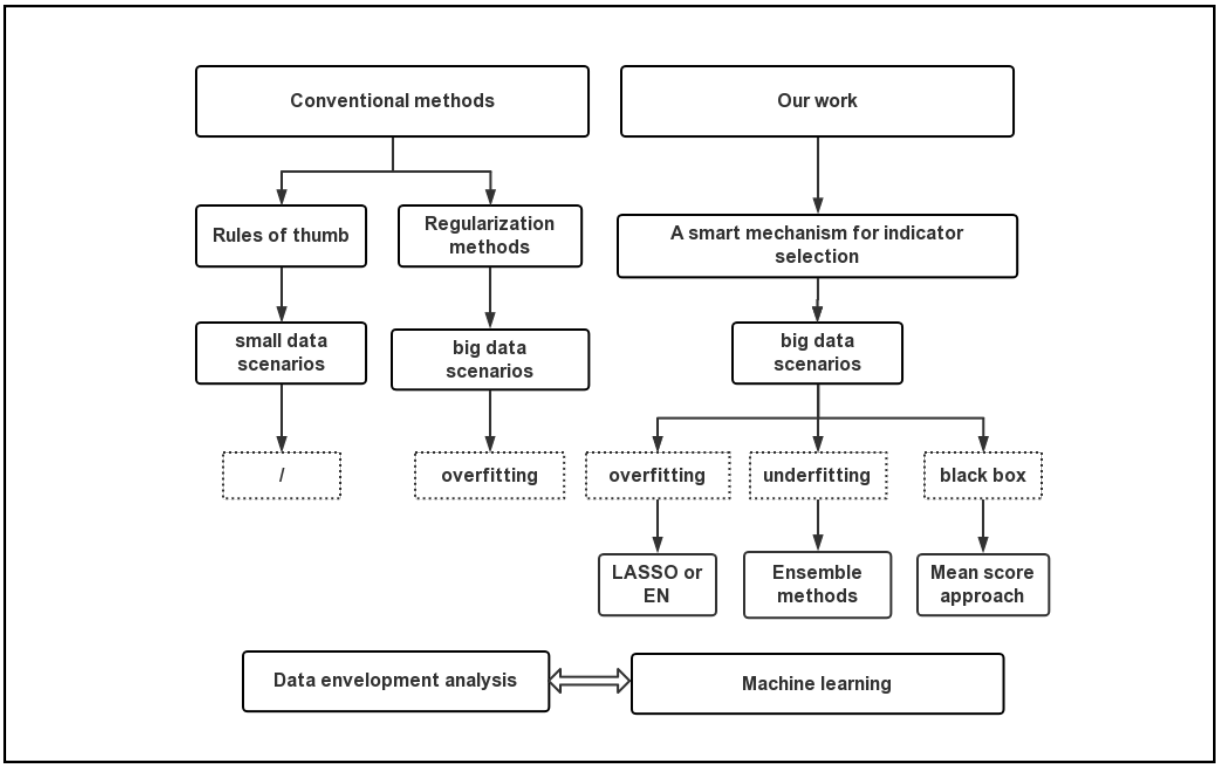
Machine learning in data envelopment analysis: A smart mechanism for indicator selection

Jie Wu, and Yumeng Wu ✉

School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Yumeng Wu, E-mail: wuyumeng@mail.ustc.edu.cn

Graphical abstract



The overall framework of our research.

Public summary

- The purpose of this study is to categorize big data scenarios encountered by data envelopment analysis into overfitting and underfitting scenarios, and to develop a smart indicator selection mechanism based on numerous machine learning approaches.
- This study advances the development of data envelopment analysis in the context of big data and combines it with machine learning to provide a convenient and intelligent indicator selection mechanism for scholars in the field of efficiency evaluation.
- Most machine learning approaches in indicator selection are confined to certain circumstances, according to Monte Carlo simulations, but the proposed mean score methodology performs well in both overfitting and underfitting scenarios.

Citation: Wu J, Wu Y M. Machine learning in data envelopment analysis: A smart mechanism for indicator selection. *JUSTC*, **2022**, 52(12): 5. DOI: 10.52396/JUSTC-2022-0106

JUSTC

Article

http://justc.ustc.edu.cn

Received: July 20, 2022; Accepted: September 05, 2022

Machine learning in data envelopment analysis: A smart mechanism for indicator selection

Jie Wu, and Yumeng Wu ✉

School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Yumeng Wu, E-mail: wuyumeng@mail.ustc.edu.cn

Cite This: *JUSTC*, **2022**, 52(12): 5 (8pp)

Read Online

Abstract: Indicator selection has been a compelling problem in data envelopment analysis. With the advent of the big data era, scholars are faced with more complex indicator selection situations. The boom in machine learning presents an opportunity to address this problem. However, poor quality indicators may be selected if inappropriate methods are used in overfitting or underfitting scenarios. To date, some scholars have pioneered the use of the least absolute shrinkage and selection operator to select indicators in overfitting scenarios, but researchers have not proposed classifying the big data scenarios encountered by DEA into overfitting and underfitting scenarios, nor have they attempted to develop a complete indicator selection system for both scenarios. To fill these research gaps, this study employs machine learning methods and proposes a mean score approach based on them. Our Monte Carlo simulations show that the least absolute shrinkage and selection operator dominates in overfitting scenarios but fails to select good indicators in underfitting scenarios, while the ensemble methods are superior in underfitting scenarios, and the proposed mean approach performs well in both scenarios. Based on the strengths and limitations of the different methods, a smart indicator selection mechanism is proposed to facilitate the selection of DEA indicators.

Keywords: data envelopment analysis; overfitting; underfitting; machine learning; indicator selection

CLC number: TP181 **Document code:** A

1 Introduction

Data envelopment analysis (DEA) has long been used to compare and evaluate the efficiency of a set of decision-making units (DMUs) with multiple inputs and outputs[1–3]. DEA is prone to overfitting (a large number of DMUs are evaluated as efficient) when a limited number of DMUs are employed to estimate a high-dimensional frontier[4, 5], which is the reason for a series of discussions on the selection of DEA indicators. Golany and Roll[6] asserted that the number of DMUs should be at least twice the number of inputs and outputs considered. Boussofiane et al.[7] stated that to effectively distinguish the DEA weights, the minimum number of DMUs should not be less than the product of the number of inputs and that of outputs. Bowlin[8] and Cooper et al.[9] claimed that the number of DMUs should be at least 3 times the sum of the number of inputs and outputs to give a meaningful estimate. From a practical point of view, these rule-of-thumb decisions are reasonable in the case of small data but may not work in big data scenarios[4, 5]. The boom in machine learning techniques has provided researchers with the opportunity to identify good indicators from big data with noise, but if inappropriate indicator selection methods are used, low-quality indicators may be selected. Specifically, if the number of indicators is large compared to that of DMUs, a machine learning method is prone to overfitting, as the information provided by each DMU in high-dimensional variable space is too sufficient to obtain good training results[4, 5]. In contrast, when the number of indicators is much smaller than that of DMUs, the information provided by indicators of DMUs may be too insufficient to obtain a good result, which is the case of underfitting. In information theory, overfitting depicts a scenario where a method goes beyond learning the true regularities in the data due to the inadvertent capturing of noise in the modeling process, while underfitting is proposed to describe the scenarios where a method fails to capture enough information to obtain the true patterns of the data[10]. For the sake of visualization, we introduce the scenarios based on the two application conditions in the real world. One is that the Ministry of Education of the People's Republic of China plans to use DEA to evaluate the performance of 20 universities in China, with a total of 100 indicators for 20 universities, which is a typical overfitting scenario; the other is that Quacquarelli Symonds (QS) intends to use DEA to evaluate the performance of 1000 universities worldwide to produce *QS World University Rankings*, but some of the indicators are not publicly available, so QS only obtains 20 indicators for 1000 DMUs, corresponding to an underfitting scenario. In both scenarios, indicator selection is essential. Indicator selection in overfitting scenarios has attracted the attention of several scholars. Ueda and Hoshiai[11] and Adler and Golany[12] successively used principal component analysis (PCA) to solve the indicator selection problem in DEA, but many of the weights of the indicators defining the principal components in PCA are negative[11], which can lead to counterintuitive integrated negative inputs[5]. Recently, some scholars[4, 5] have be-

DOI: 10.52396/JUSTC-2022-0106

JUSTC, **2022**, 52(12): 5

come pioneers in marrying the least absolute shrinkage and selection operator (LASSO)[13] with DEA, illustrating that LASSO can avoid overfitting challenges and excels in identifying good indicators, but if the relationship between outputs and inputs is nonlinear, then LASSO (using a linear model in the original unit) will be misspecified[4]. In underfitting scenarios, the relationship between outputs and inputs tends to be nonlinear, so LASSO may fail to select good indicators. To our knowledge, few studies have focused on underfitting scenarios, and researchers have not proposed classifying the big data scenarios encountered by DEA into overfitting and underfitting scenarios, nor have they attempted to develop an indicator selection system for both scenarios.

To fill these research gaps, the purpose of this study is to establish a smart indicator selection mechanism for overfitting and underfitting scenarios as an alternative to conventional methods, thus helping scholars in the field of DEA select indicators in the context of big data. In this study, the feature selection system (including regularization, wrappers, filters, and ensemble methods) is used to prevent overfitting and underfitting because regularization and wrappers are associated with cross-validation[14] and grid search[15] that are committed to overcoming overfitting, while ensemble models[16, 17] and kernel functions are developed to fit nonlinear and complex data, thereby overcoming underfitting. Based on the feature selection methods, we propose a mean score methodology that is expected to perform well in both scenarios. Then, we will learn the strengths and limitations of the different methods by applying these methods and the proposed mean score methodology in different DEA scenarios to establish a smart indicator selection mechanism. To the best of our knowledge, this paper is the first in the field of DEA to propose an indicator selection process for both overfitting and underfitting scenarios. The remainder of this study unfolds as follows: The next section briefly reviews the feature selection system, and a mean score methodology based on them is proposed in Section 3. The purpose of Section 4 is to compare and analyze the performance of the indicator selection methods by Monte Carlo simulations in both scenarios and to propose a smart mechanism for DEA indicator selection. Finally, Section 5 draws significant conclusions.

2 Preliminaries: The feature selection system

With the past few decades witnessing the prosperity of machine learning methods, indicator selection has formed a complete system, which falls into four categories based on how the selection process relates to the relevant prediction task, namely, regularization, filters, ensemble, and wrappers methods[18].

To propose an approach for indicator selection that is expected to perform well in both overfitting and underfitting scenarios, we introduce the above machine learning system for indicator selection.

2.1 Regularization methods

Assuming that a DEA scenario with n DMUs (each DMU has p inputs and one output) is equivalent to a standard regression problem with n observations, where each observation has p standardized input variables $x_{ij}, i = 1,\cdots,n; j = 1,\cdots,p$ and a dependent variable $y_i, i = 1,\cdots,n$, the objective of DEA is to estimate the following production function:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \alpha,\ i = 1,\cdots,n, \tag{1}$$

where α is the intercept and β_j is the coefficient of input variable x_{ij}.

Regularization methods include LASSO (a linear model trained with the L_1 penalty), ridge regression (linear least squares with the L_2 penalty), and elastic net (linear regression with the combined L_1 and L_2 penalty). Good indicators can be selected by imposing an L_1 or L_2 penalty on the sum of coefficients and solving the following regression problem.

$$\min_{\alpha,\beta} \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \left\|\beta_j\right\|_r = \min_{\alpha,\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij} - \alpha\right)^2 + \lambda \sum_{j=1}^{p} \left\|\beta_j\right\|_r. \tag{2}$$

Here, we interpret y_i and $\hat{y}_i$ as true and predicted values, respectively, whereas λ is the tuning parameter (or penalty price) that is chosen by cross-validation. Note that r can be 1 and 2, where $r = 2$ corresponds to ridge regression, and $r = 1$ results in LASSO. Elastic net (EN) is a linear regression with a combined L_1 and L_2 penalty, which can be illustrated as follows:

$$\min_{\alpha,\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2 + \lambda\left[\delta \sum_{i=1}^{n}\sum_{j=1}^{p}\left\|\beta_{ij}\right\| + (1-\delta)\sum_{i=1}^{n}\sum_{j=1}^{p}\left\|\beta_{ij}\right\|_2\right], \tag{3}$$

where δ denotes the ratio of the L_1 penalty.

Among the regularization methods, LASSO shrinks the coefficient of certain indicators to zero so that indicators with nonzero coefficients can be selected, and scholars[4, 5] have demonstrated its powerful ability to select indicators in overfitting scenarios. However, LASSO will be misspecified if the relationship between outputs and inputs is nonlinear and thus may not select good indicators in underfitting scenarios.

2.2 Filter methods

Filter methods (mutual information[19] and Pearson correlation coefficient[20]) rank indicators without prediction models and have the fastest running time; however, they do not consider variable dependencies and evaluate each variable separately[21]. Moreover, they are not always effective for improving the generalization ability of the model, which is why filter methods are often used as an informative indicator selection method, as is done in this study.

Mutual information (MI) measures how much the information of one random indicator is communicated with another[22]. Here, high MI means a large reduction in uncertainty, while low MI indicates a small reduction. In this study, MIC[23] is used to measure the degree of association between two variables because it has higher accuracy than MI.

2.3 Ensemble methods

Gradient boosting regression (GBR) and random forest (RF)

DOI: 10.52396/JUSTC-2022-0106
JUSTC, 2022, 52(12): 5

are two powerful ensemble methods that reduce the number of indicators by solving predictive models[18] and can be solved by popular programming software.

Random forest (RF) creates forests of randomized decision trees to handle high-dimensional data, which is why it is suitable for underfitting scenarios. In RF, the depth in the tree where the indicators are used as decision nodes is employed. The choice of depth is vital since it ultimately determines the number of samples of the input dataset whose predictions will depend on the given indicators. The indicator selection process can be well performed as the variance is sufficiently reduced.

Gradient boosting (GB) is capable of discovering complex relationships between the indicators and the dependent variable for underfitting scenarios. In the GB process, we scale up the complexity with each iteration. At each stage, the error of each model (i.e., the gradient of the loss function) is computed. In the next stage, the new model makes up for the shortcomings of the previous weak model[17].

2.4 Wrapper methods: Recursive feature elimination

Wrapper methods[24] evaluate the performance of variable subsets on preselected predictors and have the advantages of better generalization and robust interaction with the classifier used for indicator selection[21]. Recursive feature elimination (RFE) is a greedy algorithm for finding the optimal subset of indicators. It investigates and determines the optimal variable subset by repeatedly removing irrelevant indicators. Under RFE, models (regression model or machine learning model) are constructed repeatedly. In this study, random forest-based recursive feature elimination (RF-RFE)[16] and support vector regression-based recursive feature elimination (SVR-RFE)[25] are used. Each time the chosen model is built, the best or worst performing variable based on variable coefficients is selected and set aside. The remaining indicators are then employed to repeat the same process, which is stopped until there are no more indicators to create a model. Ultimately, all indicators are sorted according to the order in which they were eliminated, resulting in a subset of the best performing indicators. However, the stability of RFE depends on which model is used at the bottom of the iterative process. If a model without regularization is used, then RFE is unstable.

3 Methodology

3.1 A mean score methodology

To build on the strengths and avoid the weaknesses of the above machine learning methods, a mean score methodology (Fig. 1) is proposed in this section. It consists of four types of methods:

Type Ⅰ. Linear methods and regularization: ① Linear regression (Linear); ② LASSO; ③ Ridge regression (Ridge); ④ EN (the ratio of L_1 penalty is 0.7).

Type Ⅱ. Filter methods: ⑤ Pearson coefficient (Pearson); ⑥ MIC.

Type Ⅲ. Wrapper methods: ⑦ RF-RFE; ⑧SVR-RFE.

Type Ⅳ. Ensemble methods: ⑨ RF; ⑩ GBR.

The process of the mean score methodology is as follows:

Step 1: In each run, the score for each indicator is obtained based on the indicator importance (coefficient) under each method. In line with Chen et al.[4], the tuning parameters are chosen automatically for each method (using cross-validation);

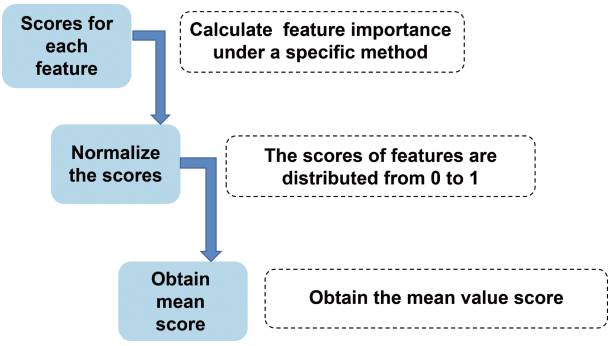


Fig. 1. Flow chart for the mean score methodology.

Step 2: Normalize the indicator score under each method so that each indicator score is between 0 (lowest-ranked indicator) and 1 (top-ranked indicator);

Step 3: The scores obtained from all methods are averaged to obtain the mean importance score of an indicator, and the indicators are ranked according to the scores.

3.2 A three-step approach for selecting indicators in the DEA context

To apply the proposed mean score methodology and the above methods to DEA indicator selection, we propose a three-step approach:

① Obtain indicator importance score: Obtain the indicator importance score under each method by running the program.

② Remove indicators: After obtaining the importance score under each method, the input indicators are removed based on the importance scores (since the scores are obtained from their correlation with the output, we consider indicators with scores less than 0.2 to be insignificant, and all indicators with scores less than 0.2 will be eliminated).

③ Run DEA: Based on the indicators selected in the previous step, the efficiency of each DMU is estimated by the following output-oriented variable return-to-scale DEA estimator.

$$\begin{aligned} &\min \theta_o \\ &\text{s.t.} \sum_{k\in K}\lambda_k x_{ki} \leqslant x_{oi}, \forall i \in I; \\ &\sum_{k\in K}\lambda_k y_{kj} \geqslant \theta_o y_{oj}, \forall j \in J; \\ &\sum_{k\in K}\lambda_k = 1,\ \lambda_k \geqslant 0, \forall k \in K; \end{aligned} \tag{4}$$

where index $o \in K$ indicates a specific DMU (which is an alias of index k), and the decision variables θ and λ_k are the efficiency estimator and the intensity multiplier for a linear combination of DMUs, respectively[5].

4 Monte Carlo results and discussion

The purpose of this section is first to see if the above methods in Section 3 can obtain accurate indicator importance scores to help identify truly relevant indicators and discard irrelevant ones, and then to illustrate the performance of these

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5

methods by the proposed three-step approach in different DEA scenarios.

4.1 Indicator identification ability

In this section, we compare the performance of all methods in Section 3 in overfitting and underfitting scenarios to explore their indicator identification capabilities.

4.1.1 Data generating process

Real-world data are complex, with not only linear but also nonlinear relationships between indicators and the output. Therefore, instead of using the DGP of Chen et al.[4], we consider complex scenarios where both linear and nonlinear relationships exist between indicators (inputs) and the dependent variable. In addition, some interfering variables associated with the truly relevant indicators are also considered. The DGP used in this section is based on Friedman regression datasets[26, 27]. The true output y^{T} and the observed output y are generated according to the following equations:

$$y^{\mathrm{T}} = 10\sin(\pi \cdot x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5, \quad (5)$$

$$y = 10\sin(\pi \cdot x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon, \quad (6)$$

where the first 5 indicators $x_1, \cdots, x_5$ have a real effect on the production function, the last 4 indicators are related to the first 4 true indicators $x_1, \cdots, x_4$, which are generated by $f(x) = x + N(0, \rho)$ (the correlation coefficient $\rho = 0.025$), and the other indicators are independent of the dependent variable. Note that the last 4 indicators were added by us, considering that some indicators may be interrelated in real life (the deviation $\varepsilon \sim N(0,1)$).

Assume that the researcher does not know which indicators are truly correlated with the dependent variable, and we use this DGP to generate 30 datasets (replications) in the following scenarios:

① Overfitting scenario Ⅰ (overfitting scenario in a small dataset): The dataset includes 100 indicators (5 true inputs, 4 correlated inputs, and 91 noisy inputs) and one output, but only 30 DMUs. In this scenario, machine learning methods tend to overfit ($p = 100, n = 30$; hereafter, n denotes the total number of DMUs, and p means the total number of indicators).

② Overfitting scenario Ⅱ (overfitting scenario in a large dataset): There are 100 DMUs in total. The large dataset contains 1000 indicators (5 true inputs, 4 correlated inputs, and 991 noisy inputs) and one output ($p = 1000, n = 100$).

③ Underfitting scenario Ⅰ (underfitting scenario in a small dataset): The dataset contains 5 true inputs, 4 correlated inputs, 11 noisy inputs, and 1 output (the dataset includes 20 potential inputs and gives no clue about the correct model specification). In this scenario, machine learning methods tend to underfit since the number of DMUs is only 100, which is too small relative to the number of indicators ($p = 20, n = 100$).

④ Underfitting scenario Ⅱ (underfitting scenario in a large dataset): 5 true inputs, 4 correlated inputs, 91 noisy inputs, and one output, and the number of DMUs is 100 ($p = 100, n = 1000$).

4.1.2 Comparison and analysis

First, it is necessary to explain our criteria for classifying large and small datasets: if $n \geq 1000$ or $p \geq 1000$, then the dataset is considered a large dataset; otherwise, it is a small dataset.

Next, we will describe our process for performing Monte Carlo simulations:

① Our program uses cross-validation to automatically choose the tuning parameters for each method.

② Each method will give every input indicator an importance score, which indicates the degree of association between this indicator and the dependent variable.

③ The final score will be normalized to the interval $[0,1]$ for comparison purposes.

④ Mean squared error (MSE) and mean absolute error (MAE) are used to analyze and compare the indicator identification ability of the above methods:

$$\mathrm{MSE}_m = \frac{1}{n}\sum_{i=1}^{n}(\theta_{im}^{\mathrm{T}} - \theta_{im}^{*})^2, \ \forall m = 1, \cdots, M, \quad (7)$$

$$\mathrm{MAE}_m = \frac{1}{n}\sum_{i=1}^{n}\left|\theta_{im}^{\mathrm{T}} - \theta_{im}^{*}\right|, \ \forall m = 1, \cdots, M, \quad (8)$$

where θ_{im}^{T} is the expected indicator importance score (theoretically, the expected importance score of the true indicator should be 1, while that of others should be shrunk to 0), and θ_{im}^{*} is the indicator importance score estimated by an indicator selection method. After M Monte Carlo experiments, the average MSE (AMSE) and the average MAE (AMAE) are as follows:

$$\mathrm{AMSE}_m = \frac{1}{M}\sum_{m=1}^{M}\mathrm{MSE}_m = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}(\theta_{im}^{\mathrm{T}} - \theta_{im}^{*})^2, \quad (9)$$

$$\mathrm{AMAE}_m = \frac{1}{M}\sum_{m=1}^{M}\mathrm{MAE}_m = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}\left|\theta_{im}^{\mathrm{T}} - \theta_{im}^{*}\right|. \quad (10)$$

To facilitate comparison, underfitting and overfitting scenarios are discussed separately.

First, the above methods for identifying the five true inputs in underfitting scenarios (both large and small datasets) are considered. Fig. 2 shows the box plots for the MSEs in 30 datasets, from which we can see that the difference in performance for these methods is fairly large:

① For both large and small datasets, GBR is the best among all methods, as the MSE under GBR is much smaller than that under other methods, especially when we focus on the median (the red line in Fig. 2); the second-best method is the proposed mean score methodology.

② SVR-RFE and RF-RFE have the worst performance, which shows that RFE is not suitable for either large or small datasets in underfitting scenarios. Following them, the MSE of LASSO is larger than that of others, indicating LASSO's inability to identify true input indicators in underfitting scenarios. In addition, LASSO has the widest data distribution, meaning that LASSO is rather unstable and may select the

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5

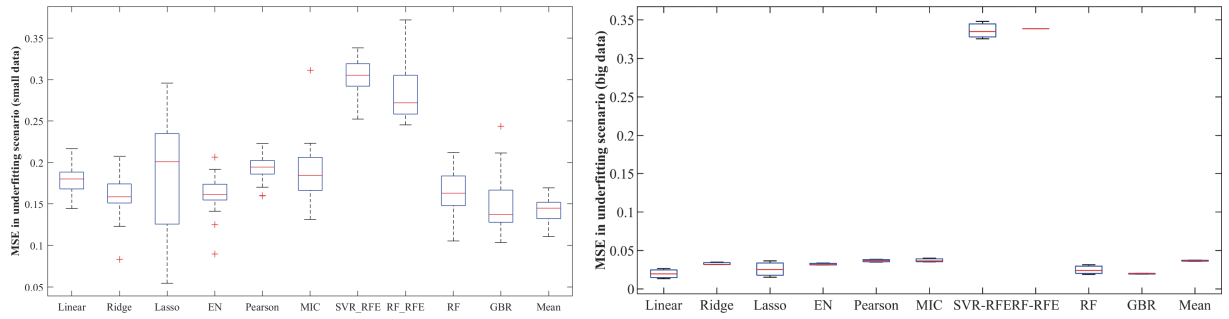


Fig. 2. Comparison of indicator identification ability of methods in underfitting scenarios.

wrong indicators instead of the true ones in this scenario.

③ Among all the regularization methods for both large and small datasets, LASSO is the worst performing method, but EN and Ridge perform quite well and are more stable. The reason is that there are correlations between indicators in the DGP (known as multicollinearity in statistics that LASSO is not capable of handling), but EN and Ridge are different, because they both use L_2 regularization, which is more stable than L_1 regularization. Therefore, LASSO is not recommended in underfitting scenarios.

Second, the indicator identification ability of each method in overfitting scenarios is compared. Fig. 3 reports the results and shows that:

① The worst performances still belong to SVR-RFE and RF-RFE for both large and small datasets.

② The "top method" turns into LASSO in overfitting scenarios whether it is a big dataset or a small one. Close behind LASSO, ensemble methods (GBR and RF) and the mean score methodology perform well.

③ For linear regression and regularization methods, the performances of EN, Ridge, and linear regression are good and stable but not as outstanding as LASSO in these two overfitting scenarios. Hence, we suggest using LASSO in preference in overfitting scenarios.

4.1.3 *Insights*

The following insights are derived based on the above simulation experiments:

① In overfitting scenarios, linear methods (LASSO, Ridge, EN, and linear regression) are preferred. The experimental results that LASSO outperforms other machine learning methods provide experimental evidence for previous literature[4, 5] on the application of LASSO to DEA.

② In underfitting scenarios, LASSO fails to identify good indicators, while ensemble methods (GBR and RF) perform well, which supports the claim in the introduction that LASSO is not suitable for the underfitting scenarios.

③ The proposed mean score methodology can accurately identify the true indicators in both scenarios, which is ideal for "black-box" environments where researchers do not distinguish between underfitting or overfitting scenarios.

4.2 Indicator selection performance in DEA scenarios

The purpose of this section is to compare the indicator selection performance of the four methods in Section 3.1 in different scenarios by comparing the true DEA efficiency scores (without noise) with the estimated values after indicator selection through the proposed three-step approach for selecting indicators in Section 3.2.

4.2.1 *Data generating process*

The DGP of Lee and Cai[5] is used to generate the production function with 100 inputs and 1 output.

$$y_k^T = \prod_{i\in I} x_{ki}^{\left(\frac{1}{|I|+1}\right)}, \forall k = 1, \cdots, n, \tag{11}$$

$$y_k^T = \prod_{i\in I} x_{ki}^{\left(\frac{1}{|I|+1}\right)} \times \mathrm{e}^{-\mu_k}, \forall k = 1, \cdots, n, \tag{12}$$

where x_{ki} represents the input i of the DMU k, and y_{kj} denotes the output. The inputs are generated by a uniform distribution of the interval $(10, 20)$, and the inefficiency term μ_k is half-normal with a mean of 0 and variance of 0.7. Here, y_k^T is the output of a true frontier, and the observed output y_k represents an output affected by inefficiency.

Three sample sizes with 30 replications are considered in our Monte Carlo simulations:

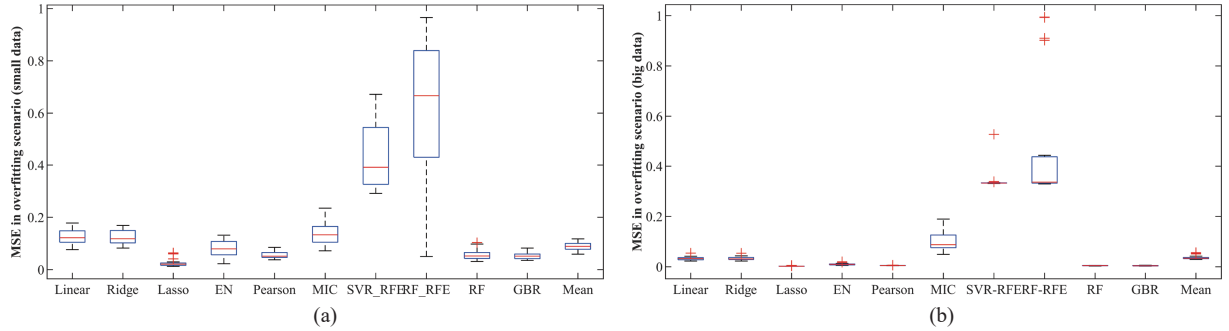


Fig. 3. Comparison of indicator identification ability of methods in overfitting scenarios.

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5

① Overfitting scenario: the DGP has 30 DMUs ($n = 30$), and each DMU contains 101 dimensions ($p = 100$, 100 inputs and 1 output);

② Underfitting scenario in a large dataset: this DGP contains 1000 DMUs ($n = 1000$), each with 101 dimensions ($p = 100$, 100 inputs and 1 output).

In addition, to make the experimental results more convincing, we add another set of data to describe underfitting scenarios:

③ Underfitting scenario in a small dataset: 100 DMUs ($n = 100$), each with 21 dimensions ($p = 20$, 20 inputs and 1 output).

4.2.2 Comparison and analysis

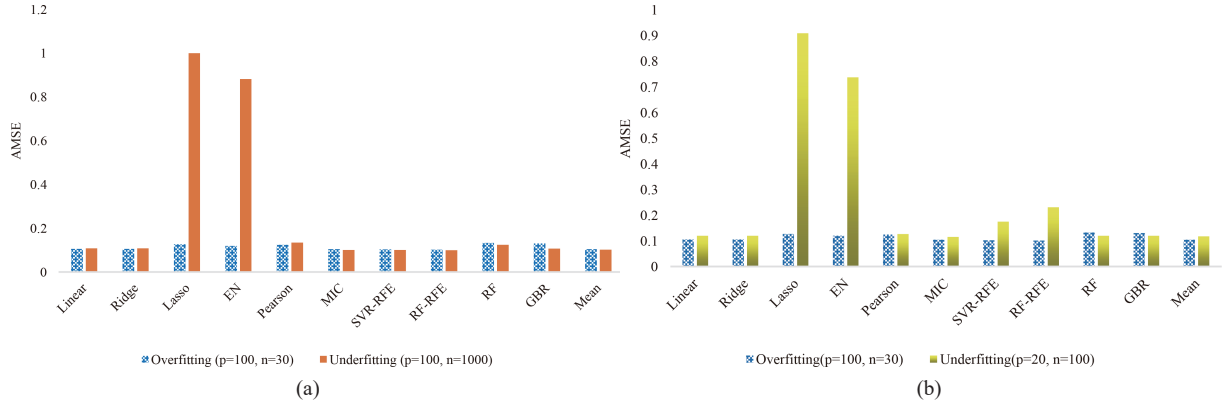
After running the proposed three-step approach, MSE, MAE, AMSE, and AMAE between each true frontier and estimated frontier under each method can be obtained by Eqs. (7)–(10). Note that θ_{im}^{T} and θ_{im}^{*} indicate the true DEA efficiency score and the estimated score after indicator selection, respectively. Table 1 and Fig. 4 report the results.

First, the change in the AMSE of each method is analyzed from overfitting to underfitting scenarios of Table 1. Theoretically, if we fix $p = 100$, then when n changes from 30 (an overfitting scenario) to 1000 (an underfitting scenario), the linear model-based methods tend to fail to select good indicators, resulting in a large deviation between the estimated value and the true value of efficiency, as evidenced in Table 1. From Table 1, linear regression, regularization methods (LASSO, Ridge, and EN), and Pearson show increases in AMSE, with the most pronounced rise in LASSO and EN (see Table 1 and Fig. 4a, LASSO increases from 0.127 to 1.0000 and EN rises from 0.120 to 0.882). The main reason may be that LASSO and EN are not suitable for fitting linear data rather than nonlinear data (data relationships tend to be linear in overfitting scenarios but nonlinear in underfitting scenarios). MIC, SVR-RFE, RF-RFE, RF, and GBR, which can fit complex data and identify nonlinear relationships, are considered in underfitting scenarios, and the results in Table 1 illustrate the superiority of these methods. The AMSE of these methods decreases from the overfitting scenario ($p = 100, n = 30$) to the underfitting scenario ($p = 100$, $n = 1000$), and our mean score methodology also performs well (AMSE decreases from 0.105 to 0.103).

The underfitting scenario with a small dataset ($p = 20, n = 100$) is considered for comparison with the overfitting scenario ($p = 100, n = 30$), and Table 1 and Fig. 4b demonstrate the results. From the overfitting scenario ($p = 100, n = 30$) to this underfitting scenario ($p = 20$, $n = 100$), LASSO and EN show significant increases in AMSE, while both RF and GBR show decreases. These findings are consistent with the above analysis when the overfitting scenario ($p = 100, n = 30$) is transferred to the underfitting scenario in the large dataset ($p = 100, n = 1000$). A differ-

Table 1. Comparison of methods in overfitting and underfitting scenarios.

	Scenarios	Type Ⅰ Linear regression and regularization				Type Ⅱ		Type Ⅲ		Type Ⅳ		Mean
		Linear	Ridge	LASSO	EN	Pearson	MIC	SVR-RFE	RF-RFE	RF	GBR	
AMSE	Overfitting (p=100, n=30)	0.106	0.106	0.127	0.120	0.125	0.105	0.103	0.102	0.133	0.131	0.105
	Underfitting in small dataset (p=20, n=100)	0.120	0.120	0.909	0.737	0.127	0.116	0.175	0.232	0.121	0.120	0.118
	Underfitting in big dataset (p=100, n=1000)	0.108	0.108	1.000	0.882	0.134	0.101	0.101	0.100	0.125	0.107	0.102
AMAE	Overfitting (p=100, n=30)	0.213	0.212	0.268	0.244	0.247	0.210	0.206	0.205	0.268	0.278	0.210
	Underfitting in small dataset (p=20, n=100)	0.249	0.249	0.924	0.780	0.270	0.242	0.291	0.339	0.252	0.249	0.246
	Underfitting in big dataset (p=100, n=1000)	0.232	0.232	1.000	0.757	0.288	0.216	0.218	0.217	0.275	0.234	0.221



(a) (b)

Fig. 4. AMSE of methods between the overfitting and underfitting scenarios.

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5

ent conclusion is that the AMSE increases significantly for SVR-RFE and RF-RFE, indicating that the RFE-based approach is unstable. The reason is that RFE is dependent on its underlying method, and if the underlying method is unstable, RFE will hardly perform well.

In Fig. 4, except for LASSO and EN, there is no significant difference in the performance of most methods between underfitting and overfitting scenarios, meaning that in underfitting scenarios, LASSO and EN (Type Ⅰ) tend to fail in indicator selection, while MIC (Type Ⅱ), RF (Type Ⅳ), GBR (Type Ⅳ), and the mean score methodology are all good performers.

Then, the performance of each method by scenario is analyzed. In the overfitting scenario of Table 1, the AMSE of MIC, the RFE-based methods (RF-RFE and SVR-RFE) and the mean score methodology are all excellent performers, which is the same conclusion as in Section 4.1. Not in line with the findings in Section 4.1, the performance of LASSO + DEA is moderate in the overfitting scenario and is not superior to other machine learning methods. By observing LASSO's scoring of individual indicators, we find that the reason for LASSO's mediocre performance is that LASSO gives most indicators a score of 0 (LASSO may be affected by the results of automatic parameter tuning), meaning that LASSO removes most indicators, even those of high importance. The analysis based on AMSE may be biased, as the mean value tends to be affected by outliers, so the results with 30 replications are plotted in Fig. 5. From Fig. 5, results similar to Table 1 can be obtained: the two methods based on RFE, MIC, and mean score methodology perform well, and the performance of Ridge is comparable to that of linear regression; the rest of the methods performed poorly, with RF and GBR being the worst performers, which implies that we should not easily use RF or GBR (Type IV) for indicator selection in overfitting scenarios.

Next, the underfitting scenario with a large dataset ($p = 100, n = 1000$) is analyzed. From Table 1, LASSO and EN are the two worst-performing methods, while there is little difference in the performance of the other good-performing methods (see Fig. 4a). The results with 30 replications are plotted in Fig. 6a. From Fig. 6a, the performance of LASSO and EN is still the worst. To compare the performance of the well-performing methods, we remove LASSO, EN, and the two RFE-based methods (as LASSO and EN have too large MSEs, and SVR-RFE and RF-RFE are not stable) to obtain Fig. 6b. If we use linear regression as the benchmark, then MIC performs the best, followed by the mean score methodology and Ridge, and finally GBR.

The results with 30 replications in the underfitting scenario of a small dataset ($p = 20, n = 100$) are plotted in Fig. 7a, which also shows that LASSO and EN are not suitable for underfitting scenarios. Similarly, we remove LASSO, EN, and the RFE-based methods to obtain Fig. 7b, from which we can draw the same conclusions as the underfitting scenario in a large dataset: MIC performed the best, followed by mean

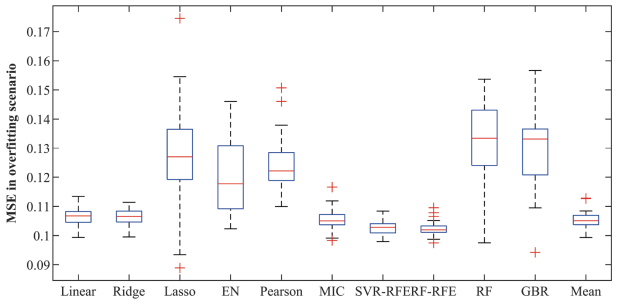


Fig. 5. MSE of methods in the overfitting scenario with 30 replications.

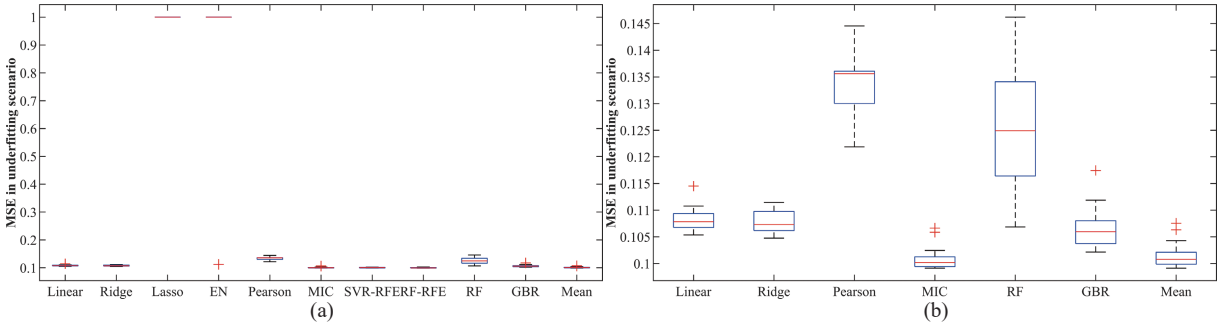


Fig. 6. MSE of methods in underfitting scenario I with 30 replications.

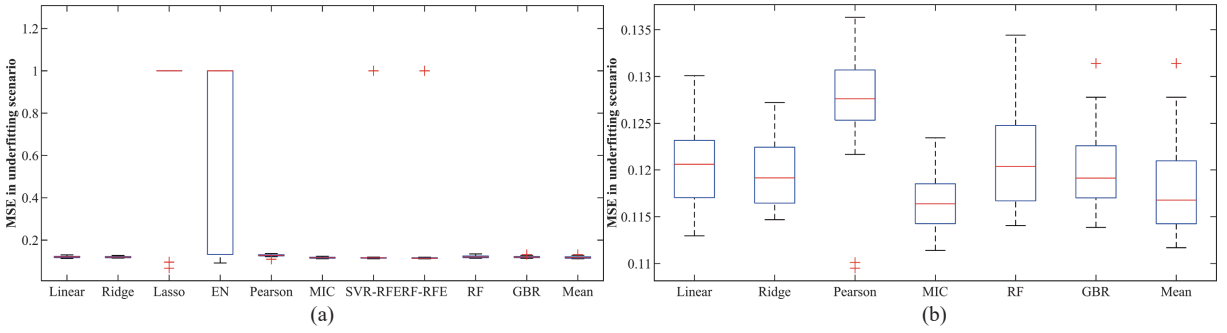


Fig. 7. MSE of methods in underfitting scenario II with 30 replications.

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5

score methodology and ridge, and then GBR.

4.2.3 Insights

In overfitting scenarios, most indicator selection methods tend to yield good results, except for ensemble methods (GBR and RF) that are prone to overfitting. Considering the nature of LASSO to produce sparse solutions and its overwhelming advantage in indicator identification, LASSO is the most recommended method in overfitting scenarios, as demonstrated in previous studies[4, 5].

In underfitting scenarios, MIC is the best choice for the univariate case (MIC identifies both nonlinear and linear relationships, and it does not require tuning parameters, leading to small computational difficulty). In the multivariate scenarios, MIC is no longer applicable, while the ensemble method (GBR) is the recommended method.

The mean score methodology proposed works well in both scenarios, which enlightens us that it may be a good choice if indicator selection is a black-box process (it is difficult to distinguish between overfitting and underfitting scenarios).

4.3 A smart indicator selection mechanism for DEA

Given the different dimensional relationships between DMUs and indicators (different DEA scenarios), indicator selection methods may encounter overfitting or underfitting. To avoid complex and time-consuming indicator engineering, it is urgent to build a process that is superior in different scenarios. Therefore, we propose a smart indicator selection mechanism to address the problem (Fig. 8).

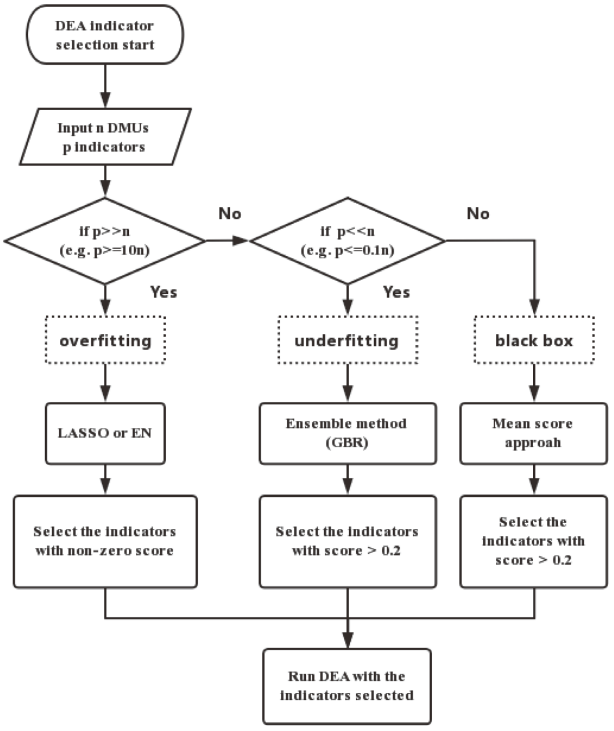


Fig. 8. A smart indicator selection mechanism for DEA.

Step 1: Input n DMUs and p indicators.

Step 2: Identify scenarios based on the relationship between p and n.

Step 3: Select indicators with the corresponding approach. LASSO or EN is recommended in overfitting scenarios, ensemble methods are suggested in underfitting scenarios, and the mean score methodology is used in the black-box scenarios.

Step 4: Run DEA with the indicators selected in Step 3.

5 Conclusions

The purpose of this study is to categorize big data scenarios encountered by DEA into overfitting and underfitting scenarios and to develop a smart indicator selection mechanism to facilitate the process of indicator selection. The main contributions are as follows:

This study is the first to classify the DEA scenarios into overfitting and underfitting scenarios and to propose a mean score methodology based on machine learning methods, advancing the development of DEA in the context of big data and its integration with machine learning.

The results of Monte Carlo experiments show that in overfitting scenarios, LASSO is superior to other methods, providing evidence for previous literature on the application of LASSO to DEA. In underfitting scenarios, however, LASSO fails, while the ensemble methods perform well, illustrating the importance of distinguishing between the underfitting and overfitting scenarios. Notably, our mean score methodology outperforms other methods in both scenarios and can serve as a good alternative in black-box environments.

However, our study has limitations because only single output scenarios are considered. Future research should consider extending our work to DEA contexts with multiple desirable outputs and undesirable outputs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (72174053, 71921001, 71971203), the Four Batch Talent Programs of China, the Fundamental Research Funds for the Central Universities (WK2040000027), and Anhui Philosophy and Social Science Foundation (AHSKY2021D147).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Jie Wu is currently a Professor at the School of Management, University of Science and Technology of China (USTC). He received his Ph.D. degree in Management Science from the USTC in 2008. His research mainly focuses on operations research.

Yumeng Wu is currently a master student at the School of Management, University of Science and Technology of China. She received her B.S. degree from the Northwestern Polytechnical University in 2016. Her research interests focus on operations research and computer science.

(Continued on page 7-9)

DOI: 10.52396/JUSTC-2022-0106
JUSTC, **2022**, 52(12): 5